Part A

1.

a. **Your team's goal is to help clients determine whether they should invest in p2p loans. What is the final decision that you will help the client make? What is the objective, and how will you evaluate 'better' vs 'worse' decisions? What is the goal of predictive models for this? What will be the potential target variables?**

- **Final Decision:** Investor get to know whether to invest in the p2p loan or not.
- **Objective:**
  - Investor's better decision would be to invest in a fully paid loan and high actual returns.
  - Investor's worse decision would be to invest in a Charged off loan and low actual returns or negative annual returns.
- **Goal of predictive models:**
  - To predict whether the loan would be a Fully paid or a Charged off loan.
  - To estimate the actual returns or annual returns
- **Potential target variables:**
  - loan status
  - actual returns or annual returns.

b. **Look at the data attributes. How would you categorize these attributes, in broad terms, considering what they pertain to? Before doing any analyses, what do you think may be the important attributes to consider for your decision task?**

- **Categorization of attributes:**
  - Borrower's demographics
  - Borrower's finance history
  - Derived attributes based on finance history (int rate, grade, subgrade, etc.)
  - Post-approval Loan details
- **Important attributes to consider, probably:**
  - Annual income
  - dti
  - latest credit pull date
  - Months since most recent bankcard delinquency
  - Number of accounts ever 120 or more days past due
  - Ratio of total current balance to high credit/credit limit for all revolving accounts

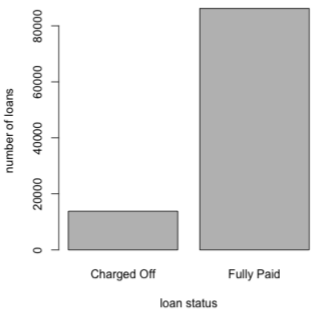
2. Data exploration:

a. some questions to consider:

i. **What is the proportion of defaults ('charged off' vs 'fully paid' loans) in the data? How does default rate vary with loan grade? Does it vary with sub-grade? And is this what you would expect, and why?**

Proportion of loans:

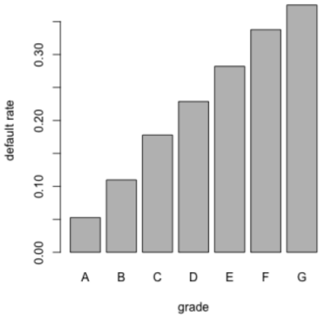
- Charged Off: 13785 (defaults)
- Fully Paid: 86215

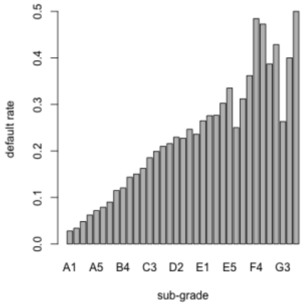


Proportion of defaults (Charged off loans) grade wise:

	A	B	C	D	E	F	G
Charged Off:	**1187**	**3723**	**4738**	**2858**	**1010**	**239**	**30**
Fully Paid:	21401	30184	21907	9635	2569	469	50

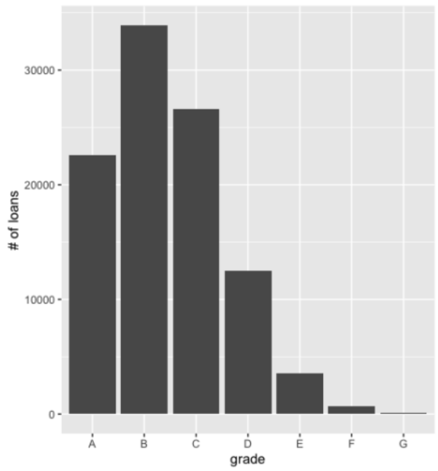
Default rate is increasing from grade ‘A’ to ‘G’ whereas default rate is not continuously increasing sub-grade wise. It has dropped in between. But overall, it is increasing. I expect the same because the chance of safest loans becoming defaults is much lesser than risky loans.

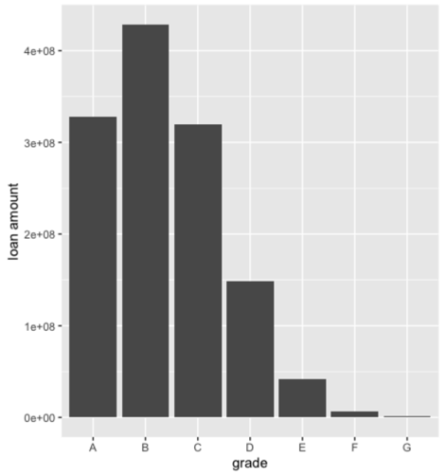


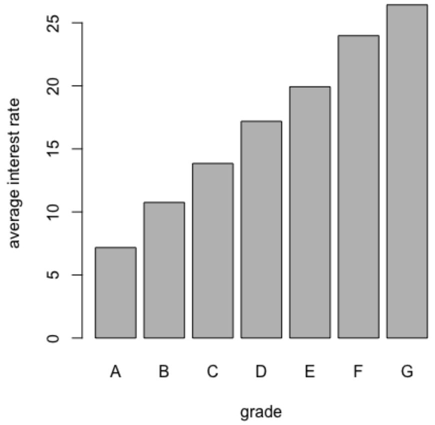


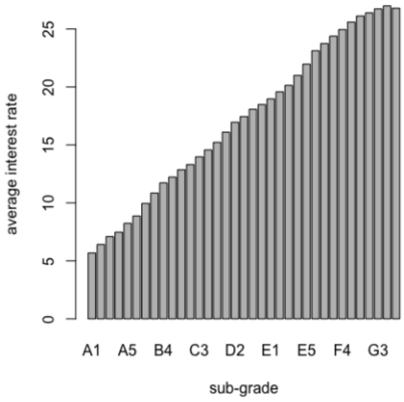
ii. How many loans are there in each grade? And do loan amounts vary by grade? Does interest rate for loans vary with grade, subgrade? Look at the average, standard-deviation, min and max of interest rate by grade and subgrade. Is this what you expect, and why?

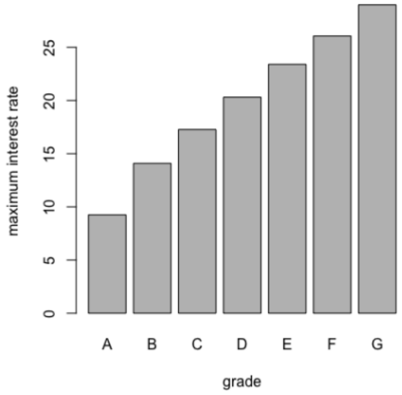
- Number of loans and Loan amount is highest for grade B loans.
- Interest rates are continuously increasing from grade A to G and also along the sub-grade wise.
- Maximum and minimum interest rates are also increasing.
- I expect the same because more risky loans should give more returns. Because there is more probability of default by risky loans.

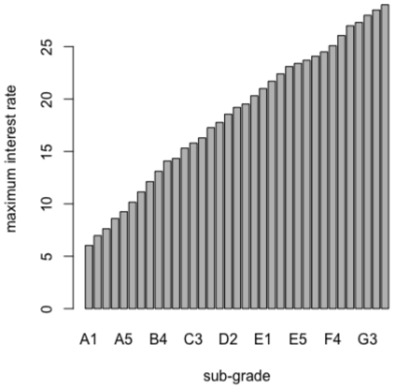






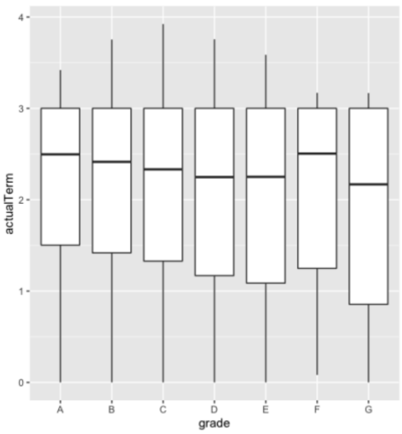






iii. For loans which are fully paid back, how does the time-to-full-payoff vary? For this, calculate the 'actual term' (issue-date to last-payment-date) for all loans. How does this actual term vary by loan grade (a boxplot can help visualize this)?

Fully paid loans are 86,215 out of 100,000 loans.



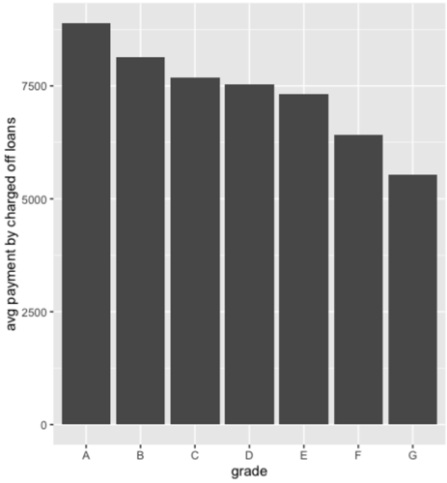
Time took to repay is not correlated much with the grade type. And also, data is not balanced. D, E, F and G grade loans are less in number. The median time took to repay doesn't have a relation with grade type.

iv. Calculate the annual return. Show how you calculate the percentage annual return. Is there any return from loans which are 'charged off'? Explain. How does return from charged - off loans vary by loan grade?

Annual return = (total repayments got - funded amount by investor)/funded amount

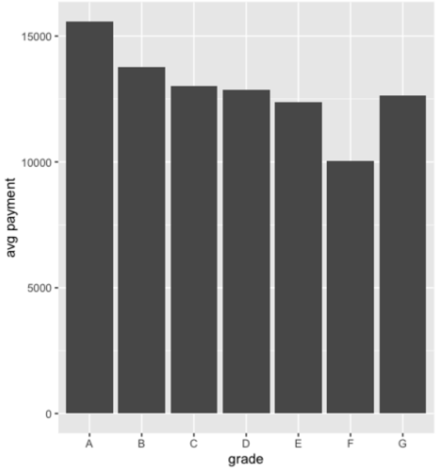
% Annual return = (total repayments got - funded amount by investor)/funded amount * (1/3) *100

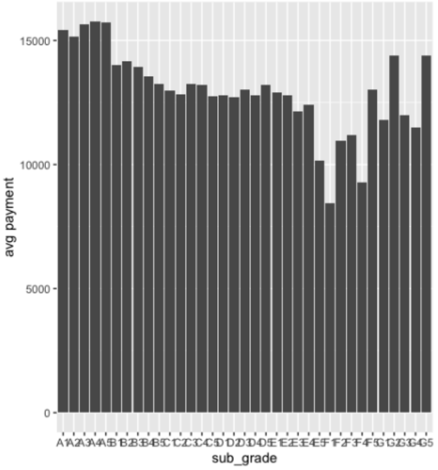
Yes, there are returns from Charged off loans. But not full payments, partial payments. Average partial payments are decreasing grade wise charged off loans.



Compare the average return values with the average interest-rate on loans – do you notice any differences, and how do you explain this? How do returns vary by grade, and by sub-grade. If you wanted to invest in loans based on this data exploration, which loans would you invest in?

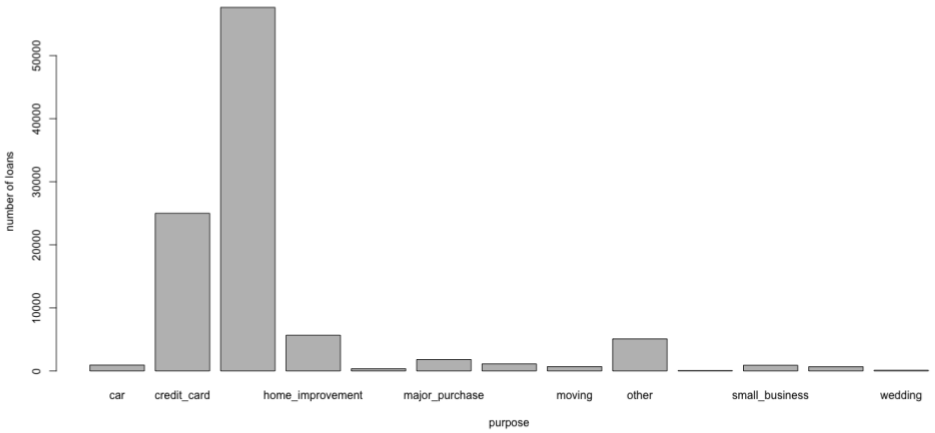
- Average return values are decreasing except at G. It might be due to small sample. Whereas average interest rate is increasing as it is dependent on the riskiness. Even though the interest rate is increasing, returns are decreasing because most of the less risky loans are fully paid.
- Same observation as grade wise return values. Decreasing along grades except at G.
- I would invest in grade 'B' loans because its annual return percentage is highest.

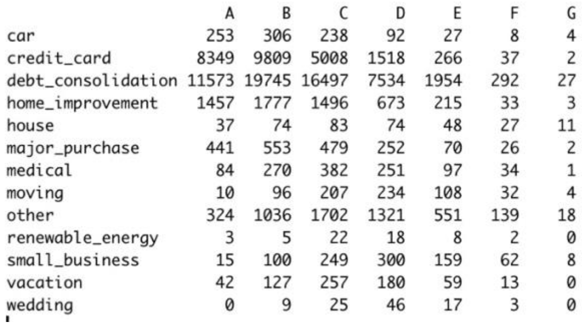




v. **What are people borrowing money for (purpose)? Examine how many loans, average amounts, etc. by purpose? Do loan amounts vary by purpose? Do defaults vary by purpose? Does loan-grade assigned by Lending Club vary by purpose?**

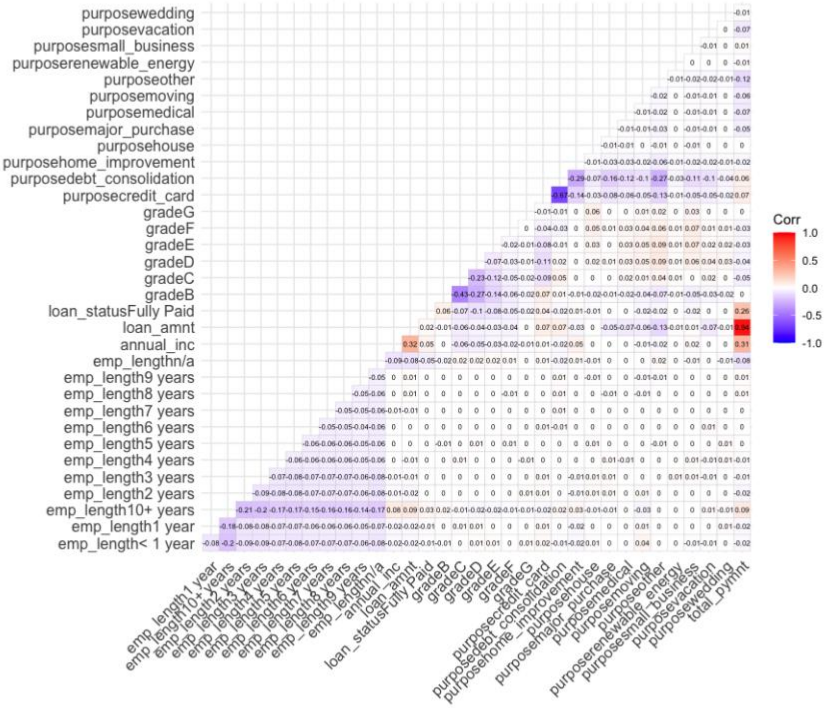
- Top 3 purposes in number of loans: Debt consolidation > credit card > other
- Top 3 purposes in average loan amounts: credit card > small business > debt consolidation.
- Defaults: debt consolidation > credit card > other. This might be due to a greater number of loans given for these purposes.
- For each purpose, ratio of number of loans by each grade is more or less similar to ratio of total number of loans along grades.




```
                       A     B     C    D    E   F  G
car                  253   306   238   92   27   8  4
credit_card         8349  9809  5008 1518  266  37  2
debt_consolidation 11573 19745 16497 7534 1954 292 27
home_improvement    1457  1777  1496  673  215  33  3
house                 37    74    83   74   48  27 11
major_purchase       441   553   479  252   70  26  2
medical               84   270   382  251   97  34  1
moving                10    96   207  234  108  32  4
other                324  1036  1702 1321  551 139 18
renewable_energy       3     5    22   18    8   2  0
small_business        15   100   249  300  159  62  8
vacation              42   127   257  180   59  13  0
wedding                0     9    25   46   17   3  0
```

vi. **Consider some borrower characteristics like employment-length, annual-income, fico-scores (low, high). How do these relate to loan attributes like, for example, loan_amount, loan_status, grade, purpose, actual return, etc.**

- Employment length is not correlated with any of the loan attributes. They are independent of employment length.
- Annual income is a bit correlated with loan amount and actual return.



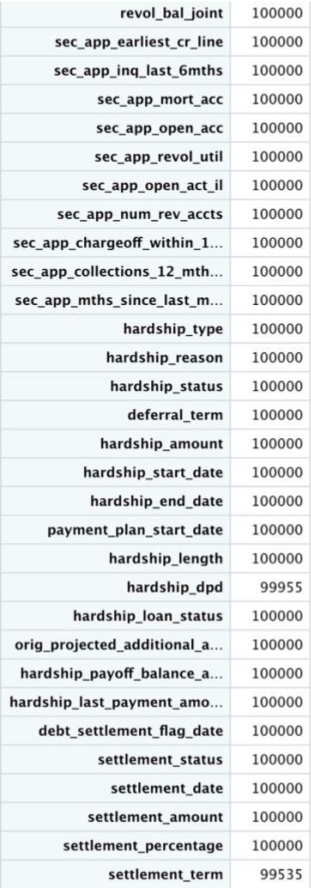
vii. **Generate some (at least 3) new derived attributes which you think may be useful for predicting default and explain what these are. For these, do an analysis as in the questions above (as reasonable based on the derived variables).**

For predicting defaults, we can check the actual returns, annual returns and actual time took to repay.

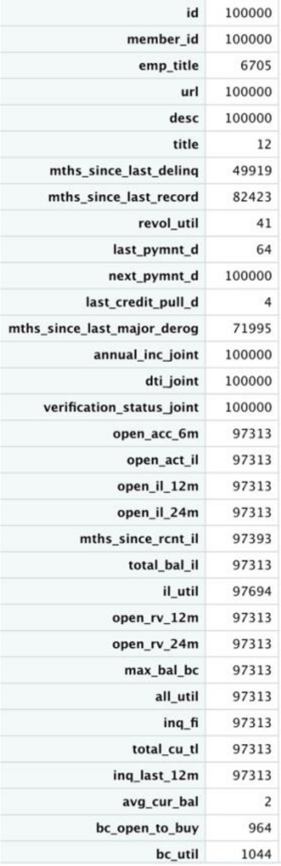
- **Actual returns**: how much returns received by the investor. If the actual returns are negative → charged off loans (default)
- **Actual term:** time taken for the repayment. If the loan is fully paid, usually it takes less than 3 years. For charged off loans, it is the maximum time period i.e. 3 years.
- **Annual returns**: for fully paid loans, it is actual returns. For charged off loans, it is low.

b. **Are there missing values? What is the proportion of missing values in different variables?**

Yes, there are 5634599 missing values in the entire data.



revol_bal_joint	100000
sec_app_earliest_cr_line	100000
sec_app_inq_last_6mths	100000
sec_app_mort_acc	100000
sec_app_open_acc	100000
sec_app_revol_util	100000
sec_app_open_act_il	100000
sec_app_num_rev_accts	100000
sec_app_chargeoff_within_1...	100000
sec_app_collections_12_mth...	100000
sec_app_mths_since_last_m...	100000
hardship_type	100000
hardship_reason	100000
hardship_status	100000
deferral_term	100000
hardship_amount	100000
hardship_start_date	100000
hardship_end_date	100000
payment_plan_start_date	100000
hardship_length	100000
hardship_dpd	99955
hardship_loan_status	100000
orig_projected_additional_a...	100000
hardship_payoff_balance_a...	100000
hardship_last_payment_amo...	100000
debt_settlement_flag_date	100000
settlement_status	100000
settlement_date	100000
settlement_amount	100000
settlement_percentage	100000
settlement_term	99535



id	100000
member_id	100000
emp_title	6705
url	100000
desc	100000
title	12
mths_since_last_delinq	49919
mths_since_last_record	82423
revol_util	41
last_pymnt_d	64
next_pymnt_d	100000
last_credit_pull_d	4
mths_since_last_major_derog	71995
annual_inc_joint	100000
dti_joint	100000
verification_status_joint	100000
open_acc_6m	97313
open_act_il	97313
open_il_12m	97313
open_il_24m	97313
mths_since_rcnt_il	97393
total_bal_il	97313
il_util	97694
open_rv_12m	97313
open_rv_24m	97313
max_bal_bc	97313
all_util	97313
inq_fi	97313
total_cu_tl	97313
inq_last_12m	97313
avg_cur_bal	2
bc_open_to_buy	964
bc_util	1044

Explain how you will handle missing values for different variables. You should consider what the variable is about, and what missing values may arise from – for example, a variable monthsSinceLastDeliquency may have no value for someone who has not yet had a delinquency; what is a sensible value to replace the missing values in this case?

- Remove the rows which have less than 10 missing values.
- In the given situation, we must replace null values with maximum value in that column because they were never delinquent.

Are there some variables you will exclude from your model due to missing values?

Yes. They are id, member_id, url, desc, next_pymnt_d, annual_inc_joint, verification_status_joint, dti_joint, open_act_il, open_il_12m, all columns related to secondary applicant, and many more.

c. Consider the potential for data leakage. You do not want to include variables in your model which may not be available when applying the model; that is, some data may not be available for new loans before they are funded. Leakage may also arise from variables in the data which may have been updated during the loan period (ie., after the loan is funded). Identify and explain which variables you will exclude from the model.

Variables such as total payment, recoveries, collection recovery fee, last payment date, etc. are not available before the loan approval or disbursement. So, if we consider these types of variables in model building, they would be highly correlated with the target variable i.e., loan status.

And variables such as grades, sub-grades, interest rates, etc. are determined while approving the loan based on applicant other past or present details. These variables need not be excluded from the model. These are dependent variables and then these will lead to the loan status more accurately.

3. **Do a univariate analysis to determine which variables (from amongst those you decide to consider for the next stage prediction task) will be individually useful for predicting the dependent variable (loan_status). For this, you need a measure of relationship between the dependent variable and each of the potential predictor variables. Given loan-status as a binary dependent variable, which measure will you use? From your analyses using this measure, which variables do you think will be useful for predicting loan_status? (Note – if certain variables on their own are highly predictive of the outcome, it is good to ask if this variable has a leakage issue).**

   Score used for univariate analysis: **AUC** (Area under Curve)

   Useful variables for predicting loan status according to primary univariate analysis are:

To consider	Not to consider
int_rate	recoveries
acc_open_past_24mths	collection_recovery_fee
annual_inc	total_rec_prncp
bc_open_to_buy	last_pymnt_amnt
tot_hi_cred_lim	total_pymnt_inv
total_bc_limit	total_pymnt
avg_cur_bal	total_rec_int
Dti	funded_amnt_inv
total_rev_hi_lim	funded_amnt
tot_cur_bal	loan_amnt
mo_sin_rcnt_tl	total_acc
mths_since_recent_inq	num_op_rev_tl
mort_acc	total_bal_ex_mort
mo_sin_rcnt_rev_tl_op	mths_since_last_major_derog
mths_since_recent_bc	mths_since_last_record
inq_last_6mths	
mo_sin_old_rev_tl_op	
bc_util	
revol_bal	
revol_util	
mo_sin_old_il_acct	
num_bc_tl	
pct_tl_nvr_dlq	
mths_since_last_delinq	
total_il_high_credit_limit	
num_il_tl	
open_acc	
num_rev_accts	
num_sats	
num_tl_120dpd_2m	

Predictive models for loan status

4.

a. **Split the data into training and validation sets. What proportions do you consider, why?**

   Proportion of training and testing data: 75:25.

   There are 100,000 records in the data set. To train the model well, ~75k records were given for training. And ~25k records are well enough for testing the model.

b. **How will you evaluate performance – which measure do you consider, and why?**

   There are two options:

   - If equal weightage is assigned to any correct prediction whether as fully paid or charged off, then **accuracy** is a good measure.
   - If cost of wrong prediction is greater for one type, then we have two options:

a. to change **threshold** and calculate accuracy.

b. to consider **precision** or **recall** measures appropriately. In the present case, cost of classifying charged off loan as a fully paid loan is lot greater than classifying fully paid loan as a charged off loan.

- For simplicity of assignment, Accuracy is considered by assuming equal cost of wrong prediction.

c. For evaluation of models, you should include confusion matrix related measures, as well as ROC analyses and lifts. Explain which performance measures you focus on, and why.

- Confusion matrix measures: accuracy, precision, recall
- Charts analyses: ROC, Lift, Gain, Profit
- **Measures to be focused for the case**: **Accuracy and ROC** (graph between TPR and FPR). Because we need the model to predict both Charged off and fully paid loans correctly. But as the cost of misprediction of charged off loan as a fully paid is more, we need to minimize FPR and increase TPR (positive class: classifying charged off loan).

5. Develop a decision tree model to predict default. Train decision tree models (use either rpart or c50). What parameters do you experiment with, and what performance do you obtain (on training and validation sets)? Clearly tabulate your results and briefly describe your findings. [If something looks too good, it may be due to leakage – make sure you address this].

Available parameters: split on (gini or info), min splits, cp, max depth, min bucket, max depth

Changed parameter	Training accuracy	Test Accuracy
split = gini; min splits = 30 (but not going up to 30)	89.08%	89.01%
Split = info; Min splits = 30	88.97%	88.91%
Split = info; Min splits = 50; cp = 0.01	88.97%	88.91%

As my dimension of training data set consists of only 38 variables (other variables were removed because of their irrelevance or variables generated after the loan process is completed), maximum number of splits were 5 which is giving the cp value of 0.01.

Identify the best tree model. Why do you consider it best? Describe this model – in terms of complexity (size).

Best tree model: split should be based on 'gini index' because both training and testing accuracy are greater than the tree which splitted based on 'info gain'. In terms of complexity, tree with info split has 5 splits and with gini split has 6 splits.

Broadly speaking, there is not much difference between both the trees.

Examine variable importance. How does this relate to your uni-variate analyses in Question 3 above?

Top 11 important variables played role in building the Decision Tree	As we have performed univariate analysis only on the numerical variables, we can only compare numeric variables from above important variables.
i. last_credit_pull_d ii. sub_grade iii. int_rate iv. grade v. bc_open_to_buy vi. total_bc_limit vii. total_rev_hi_lim viii. addr_state ix. annual_inc x. avg_cur_bal xi. total_il_high_credit_limit	**i. int_rate** ii. acc_open_past_24mths **iii. annual_inc** **iv. bc_open_to_buy** v. tot_hi_cred_lim **vi. total_bc_limit** **vii. avg_cur_bal** viii. dti **ix. total_rev_hi_lim** x. tot_cur_bal xi. mo_sin_rcnt_tl

The two above important variables are highly matched.

Briefly describe how variable importance is obtained (the process used in the decision tree learning algorithm you use. (rpart or c50).

rpart is using the CART algorithm in background. In decision trees, important variables are on top, unlike Random Forest. DT splits the data to decrease the **node impurity**. So, whichever variable could be splitted with reducing node impurity, they would be on top of the tree => important variables.

6.

a. **Develop random forest and boosted tree model (using gbm or xgb). Note the 'ranger' library and xgb can give faster computations. What parameters do you experiment with, and how does this affect performance? Describe the best random forest and boosted tree model in terms of number of trees, performance, variable importance.**

For xgb: changed eta (learning rate)

Eta	Nrounds	Accuracy_training	Accuracy_testing
	100	0.9049	0.8682
0.06	116	0.8914	0.87028
0.05	26	0.8688	0.8673
0.06	26	0.8750	0.8676

For random forest:

Mtry	**Num trees**	**OOB Error**
4	500	0.1362630
6	500	0.1352627
9	500	0.1336490
13	500	0.1334356

RF Model which is built on mtry = 13 (best): Accuracy on testing data is 0.8668.

With just 116 trees, XGBoost (87%) is giving better performance than RF (86.5%) with 500 trees.

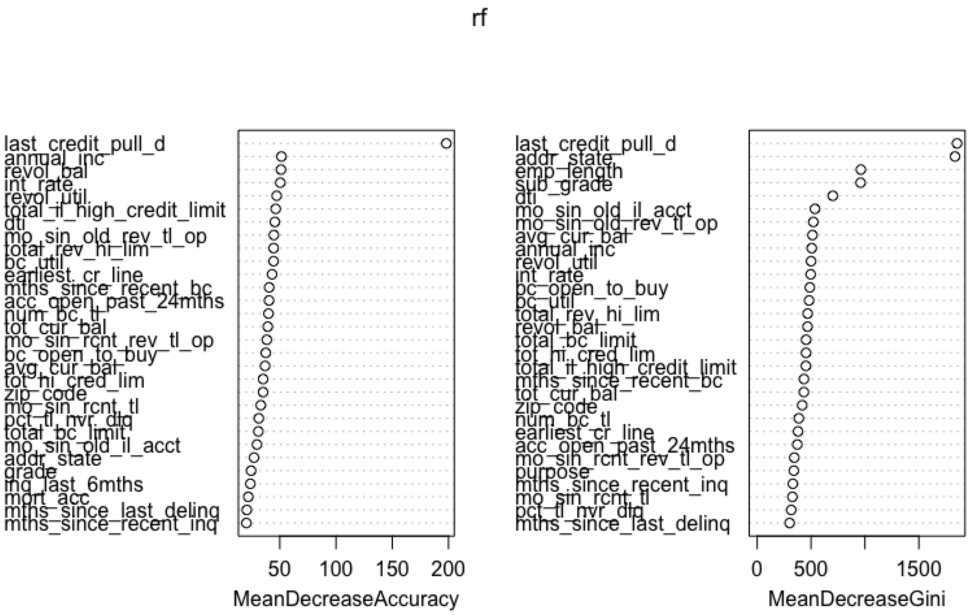


Fig.1: Important variables from Random Forest Model

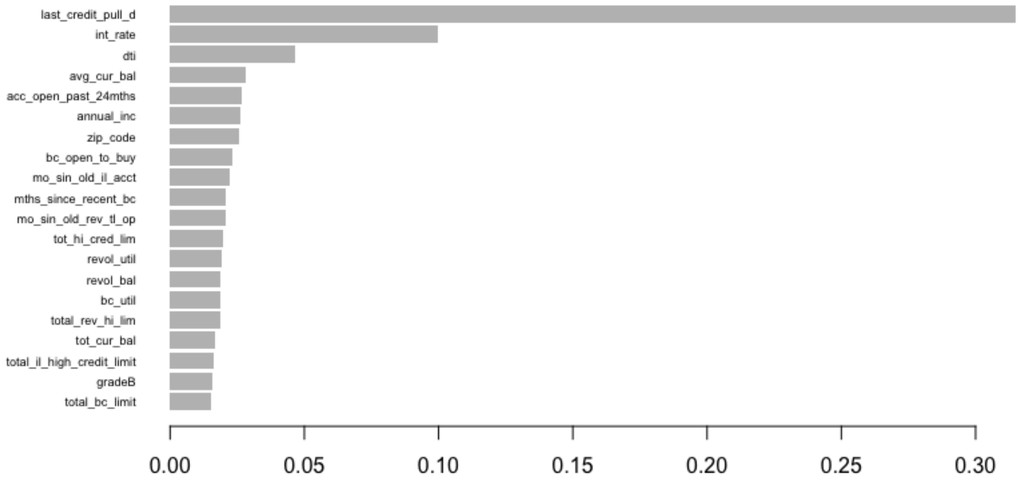


Fig.2: Important variables from XGBoost Model

b. Compare the performance of random forest, boosted tree and decision tree model from Q 5 above. Do you find the importance of variables to be different? Which model would you prefer, and why?

Usually, a single decision tree gives more variance. But, in this case, DT (~89%) is giving better accuracy on testing data than RF (86.68%) and XGBoost (87%).

Important variables: almost important variables are same in all the 3 models

DT	RF	XGB
last_credit_pull_d	last_credit_pull_d	last_credit_pull_d
sub_grade	revol_bal	int_rate
int_rate	annual_inc	dti
grade	int_rate	avg_cur_bal
bc_open_to_buy	revol_util	acc_open_past_24mnths
total_bc_limit	bc_util	annual_inc
total_rev_hi_lim	dti	zip_code
addr_state	mon_sin_old_rev_tl_op	bc_open_to_buy
annual_inc avg_cur_bal	total_il_high_credit_limit	mon_sin_old_il_acct
total_il_high_credit_limit	mnths_since_recent_bc	mnths_since_recent_bc
dti	bc_open_to_buy	mon_sin_old_rev_tl_op
	total_rev_hi_lim	tot_hi_cred_lim

So, I prefer to choose DT over RF and XGB.

7. The purpose of the model is to help make investment decisions on loans. How will you evaluate the models on this business objective?

Evaluation measure for this business objective: **Accuracy** (because model must predict the loan status accurately either fully paid or charged off)

Consider a simplified scenario - for example, that you have $100 to invest in each loan, based on the model's prediction. So, you will invest in all loans that are predicted to be 'Fully Paid'. Key questions here are: how much, on

average, can you expect to earn after 3 years from a loan that is paid off, and what is your potential loss from a loan that must be charged off?

Fully paid:

- Average interest rate for Fully paid loans: 11.71% per year (I)
- Principal = $100
- Simple interest = P*I *3 years = 100 x 0.1171 x 3 = $35.13
- Principal + Interest = $135.13

Charged off:

- Average returns on charged off loans: -11.96% per year
- Returns = funded amount (average return*3 + 1)
- Returns = 100(-0.1196*3 + 1) = $64.12
- Loss = 100 – 64.12 = $35.88

One can consider the average interest rate on loans for expected profit – is this a good estimate of your profit from a loan? For example, suppose the average int_rate in the data is 11.2%; so, after 3 years, the $100 will be worth (100 + 3*11.2) = 133.6, i.e a profit of $33.6. Now, is 11.2% a reasonable value to expect – what is the return you calculate from the data? Explain what value of profit you use.

Case 1: If the borrower has fully paid the loan before 3 years, profit should not be calculated with the average interest per year for 3 years. Here, we must calculate the profit for only the usage period (actual term).

Actual return per year =
(Total returned payments – funded amount)/funded amount * (1/actual term) *100

Profit = Funded amount * (1 + Actualreturn * actual term)

For a loan that is charged off, will the loss be the entire invested amount of $100? The data shows that such loans have do show some partial returned amount. Looking at the returned amount for charged off loans, what proportion of invested amount can you expect to recover? Is this overly optimistic? Explain which value of loss you use.

Case 2: If the loan is charged off after 3 years, the principal is not returned completely. So here we can't calculate the returned amount as Principal + Interest.

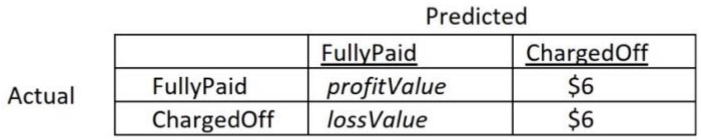
So, the optimal way for calculating loss is 'Average returns'

Avg return per year = (total returned payments – funded amount)/funded amount * (1/3 years) * 100

Loss = Funded amount * (1 – Avg return*3)

You should also consider the alternate option of investing in, say in bank CDs (certificate of deposit); let's assume that this provides an interest rate of 2%. Then, if you invest $100, you will receive $106 after 3 years (not considering reinvestments, etc.), for a profit of $6.

Considering a confusion matrix, we can then have profit/loss amounts with each cell, as follows:



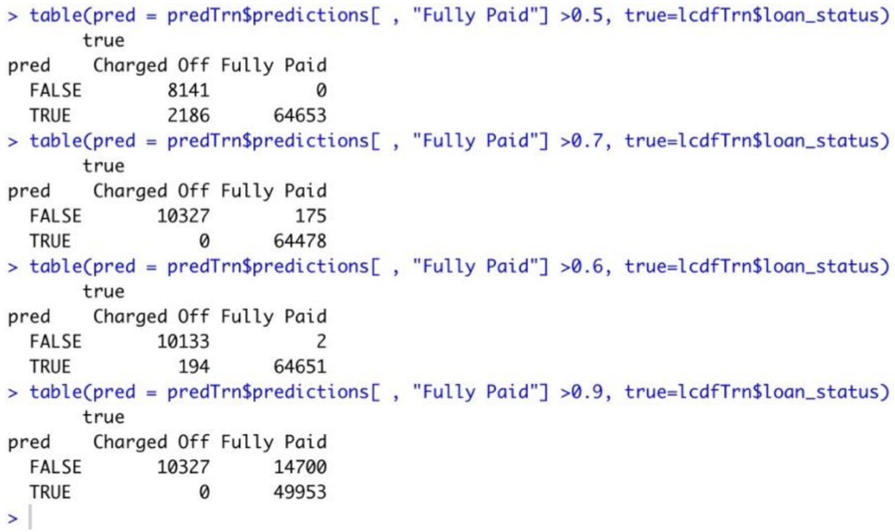
		Predicted	
		FullyPaid	ChargedOff
Actual	FullyPaid	*profitValue*	$6
	ChargedOff	*lossValue*	$6

a. Compare the performance of your models from Questions 5, 6 above based on this. Note that the confusion matrix depends on the classification threshold/cutoff you use. Which model do you think will be best, and why.

As the profit and loss values are same, the threshold would be 50%.

As all the above models are built on the very same assumption i.e. equal cost values, Decision tree model would be best for classifying more accurately than RF and XGBoost models.

b. Another approach is to directly consider how the model will be used – you can order the loans in descending order of prob(fully-paid). Then, you can consider starting with the loans which are most likely to be fully paid and go down this list till the point where overall profits begin to decline (as discussed in class). Conduct an analysis to determine what threshold/cutoff value of prob(fully-paid) you will use and what is the total profit from different models – decision tree, random forest, boosted trees. Also compare the total profits from using a model to that from investing in the safe CDs. Explain your analyses and calculations. Which model do you find to be best and why. And how does this compare with what you found to be best in part (a) above.



```
> table(pred = predTrn$predictions[ , "Fully Paid"] >0.5, true=lcdfTrn$loan_status)
       true
pred    Charged Off Fully Paid
  FALSE        8141          0
  TRUE         2186      64653
> table(pred = predTrn$predictions[ , "Fully Paid"] >0.7, true=lcdfTrn$loan_status)
       true
pred    Charged Off Fully Paid
  FALSE       10327        175
  TRUE            0      64478
> table(pred = predTrn$predictions[ , "Fully Paid"] >0.6, true=lcdfTrn$loan_status)
       true
pred    Charged Off Fully Paid
  FALSE       10133          2
  TRUE          194      64651
> table(pred = predTrn$predictions[ , "Fully Paid"] >0.9, true=lcdfTrn$loan_status)
       true
pred    Charged Off Fully Paid
  FALSE       10327      14700
  TRUE            0      49953
>
```

- **Positive class**: Fully paid
- Profit value is $6 and Loss value is $6.
- **Best Model came out: Random Forest**
- From the above thresholds, 70% i.e., **0.7** could be the best threshold. Because of 99.7% accuracy.
- Total Cumulative Profit: $ 1,874,025,126 = ~$1.87 billions

Part B: Predictive models for loans with high returns

1. Develop models to identify loans which provide the best returns.

Explain how you define returns? Does it include Lending Club's service costs?

ROI (Return on investment) = Profit/Cost i.e. (Loan amount – returns)/Loan amount. It doesn't include LC's service costs. Because service cost is not specific to the loan. Those costs are on overall.

Columns required for calculating ROI:

- **funded_amnt_inv**: Loan amount committed by an investor
- **total_pymnt_inv**: Payments received to date for portion of total amount funded by investors

Develop glm, rf, gbm (xgb) models for this. Show how you systematically experiment with different parameters to find the best models. Compare model performance.

Here, we must predict the ROI per year on each loan => so that investor can take a decision based on how much he/she have to invest (funded_amnt_inv) and how much ROI he/she is going to get.

ROI = (funded_amnt_inv - total_pymnt_inv)/ funded_amnt_inv
ROI per year = ROI / 3

We need to develop different models for "Fully paid" and "Charged Off" loans because Actual Returns for fully paid loans are different from annual returns of charged off loans. So unseen data should be subjected to predict loanStatus first and based on the predicted loan status, the unseen data then subjected to one of the models to predict either actualReturn or annRet. But for the sake of simplicity of assignment, we considered 'actual returns' to be predicted by removing 'annual returns' column.

Models:

(a) Generalized Linear Model:

- There are no parameters to be tuned in GLM Model except the 'family'. Family could be either gaussian, poisson, logit, quasi, gamma, etc.
- Our case has involved both positive and negative numbers, we have to use the family = 'gaussian'.
- Obtained the AIC: 482735 and Median Deviance Residual of 1.289

(b) Decision tree:

	minsplit	maxdepth	cp	error
(i)	17	14	0.01	0.7534755
(ii)	11	9	0.01	0.7535922
(iii)	5	10	0.01	0.7536368
(iv)	7	14	0.01	0.7536482
(v)	14	13	0.01	0.7538416

For minsplit = 17; maxdepth = 14 and cp = 0.01:

- RMSE_training: 7.688461
- RMSE_testing: 7.775561

(c) Random Forest:

Mtry	Num.trees	Node size	Rmse_training	Rmse_testing
4	200	5	3.481316	7.664623
9	300	5	3.292359	7.591041
14	400	5	3.232828	7.590556
12	350	5	3.253424	7.589385
11	320	5	3.262541	7.586748
11	320	3	3.015986	7.592762

(d) XGBoost:

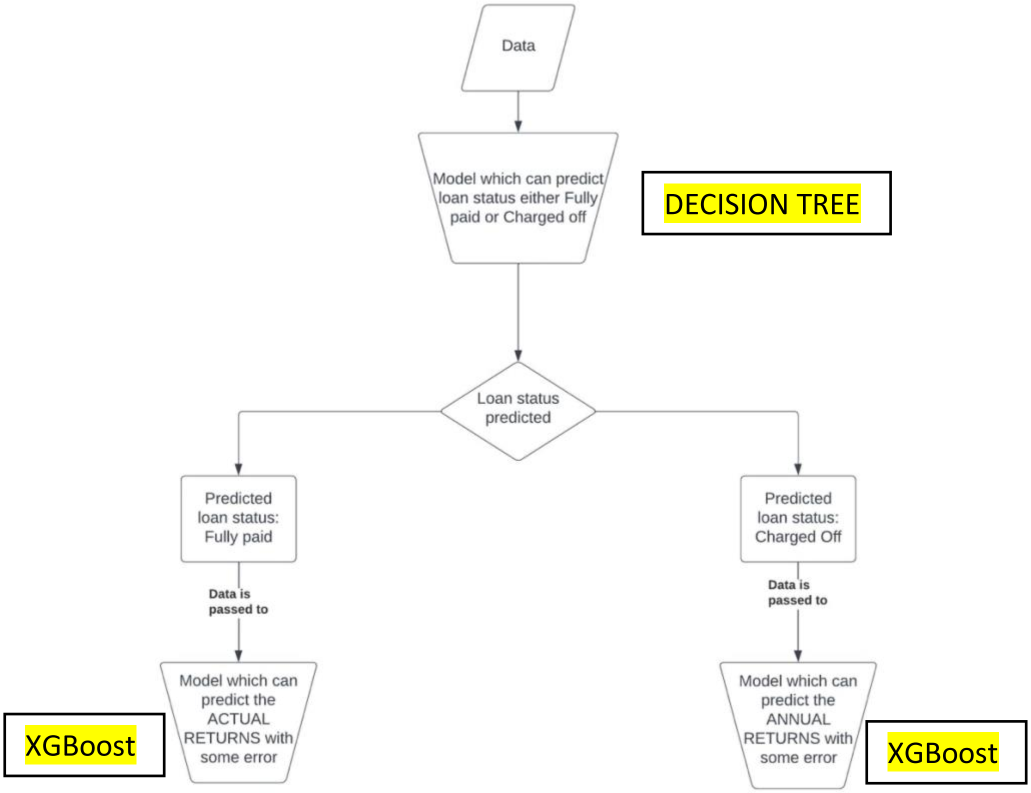
Max depth	Nrounds	Rmse_training	Rmse_testing
2	50	7.452288	7.578777
5	70	6.864012	7.550914
5	60	6.919941	7.537944
10	100	2.96749	7.937435
10	60	3.996849	7.868707
8	60	5.561327	7.704416
8	40	5.996274	7.663996
7	40	6.428261	7.598161

Best model among above models for predicting actual returns: **XGBoost**

2. **Considering results from the best model for predicting loan-status and that for predicting loan returns, how would you select loans for investment? There can be multiple approaches for combining information from the two models to make investment decisions (as discussed in class) – describe your approach and show performance. How does performance here compare with use of single (i.e for predicting loan-status, or loan returns) models?**

With single models, either predicting ‘loan status’ or ‘returns (actual or annual)’, decision given to the investor is not complete. Investor may not take decision based only on loan status or just returns.

So, after the borrower’s application for loan, with our model investor must get the complete picture and able to take a decision. My approach for it is:


Data
Model which can predict loan status either Fully paid or Charged off
DECISION TREE
Loan status predicted
Predicted loan status: Fully paid
Predicted loan status: Charged Off
Data is passed to
Data is passed to
Model which can predict the ACTUAL RETURNS with some error
XGBoost
Model which can predict the ANNUAL RETURNS with some error
XGBoost

Single models:

1. **Predicting loan status model**: ~89% accuracy
2. **Predicting returns model**: Accuracy is around 80%
3. **Pipelining both**: pipelining the data into both the models, gives the investor a better picture. The investor gets to know that:
    a. the loan would be paid off or not ~90% accurately
    b. the returns generated, if invested, ~80% accurately.

3. As seen in data summaries and your work in the first assignment, higher grade loans are less likely to default, but also carry lower interest rates; many lower grade loans are fully paid, and these can yield higher returns. Considering this, one approach to making investment decisions may be to focus on lower grade loans (C and below) and try to identify those which are likely to be paid off. Develop models from the data on lower grade loans, and check if this can provide an effective investment approach. Compare performance of models from different methods (glm, gbm, rf).

To predict whether the loan can be fully paid or charged off (ONLY FOR LOANS HAVE LOWER GRADES):

Model	Training Accuracy	Testing Accuracy
Decision tree	85.46%	85.19%
Random Forest		80.62%
XGBoost	88.33%	80.77%

It clearly shows that Decision tree is more appropriate model for predicting whether the loan would be paid off or not, even for lower grade loans and for any grade loan (which is shown in Q5.

Can this provide a useful approach for investment? Compare performance with that in Q9 above?

As we already know, grade B and grade C loans are giving more returns than grade A loans. But if the loans are charged off, it is loss for the investor. Models built in Q9 are for lower grade loans, they may give more returns, but the prediction accuracy **(~85%)** is lower than the Q5 and Q6 models **(~90%)** (which includes every loan grade).

So, if an investor invests based on the prediction of model built in Q9 may get losses.